BECREATIVE: Annotation of German Named Entities

Fabian Dreer, Eduard Saller, Patrick Elsässer, Ulrike Handelshauser, Desislava Zhekova
CIS, Ludwig Maximilian University, Munich
{dreer|sallere|elsaesser|handelshauser}@cip.ifi.lmu.de
zhekova@cis.uni-muenchen.de

Abstract

This paper presents the BECREATIVE Named Entity Recognition system and its participation at the GermEval 2014 Named Entity Recognition Shared Task (Benikova et al., 2014a). BECREATIVE uses a hybrid approach of two commonly used procedural methods, namely list-based lookups and machine learning (Naive Bayes Classification), which centers around the classifier. BECREATIVE currently reaches an F-score of 37.34 on the strict evaluation setting applied on the development set provided by GermEval.

1 Introduction

Named Entity Recognition (NER) is an important part of many natural language processing (NLP) tasks first and foremost Information Extraction (IE), but as well necessary for question-answering systems and machine translation. In general, named entities (NEs) are phrases that represent persons, organizations, locations, times, quantities, etc. (Tjong Kim Sang and De Meulder, 2003). NER is the task of locating those phrases, mostly proper names, in an unstructured text and clustering them into a predefined set of categories.

The rest of this paper is organized as follows: In section 2, related work on the topic of NER that has been carried out over the last years is presented and discussed. Following, (in section 3) we shortly present the GermEval 2014 Shared Task (Benikova et al., 2014a) in the context of which the system was developed and evaluated. The description of BECREATIVE can be then found in section 4 that is followed by its evaluation (see section 5) and conclusion (section 6).

2 Related Work

Nowadays NER has reached numerous traditional domains, such as medicine or biology, but as well a more novel domain: The internet with all its blogs and social platforms where NER tools need to be less domain specific and thus perform quite differently than on an e.g. journalistic corpus. NER was first looked into more concretely back in 1990 (Nadeau and Sekine, 2007), when the main approaches were still based on heuristics and handcrafted rules. Shortly afterwards, it was already recognized as an essential subtasks of IE. The initial purpose was to extract structured information like names of persons, locations, organizations and also numeric values like time or date from newspaper articles or specialist literature. In 1995 at MUC-6 (Grishman and Sundheim, 1996) NER was constituted to be the initial goal for the first time, so "Named Entity" became an internationally accepted term in the world of natural language processing. Prerequisite for precise NER is the segmentation of data, performed by tokenization and chunking; for example "University of Munich" is a single NE, and the token "Munich" inside its span is also a NE. Yet, detecting all NEs (Carreras et al., 2002) and classifying them by their type still is a very challenging task (Tjong Kim Sang and De Meulder, 2003). Besides NER on English texts, which is generally the language concentraiting most efforts, a small number of approaches for other languages were also carried out, such as (IREX) (Sekine and Isahara, 2000) for Japanese or as well the systems on German, Dutch or Spanish presented during the CoNLL 2002 and 2003 Shared Tasks on Language-Independent Named Entity Recognition (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003). In the IREX project and the MUC-6 NE task (muc, 1995), new categories, such as *artifact*, *geographical* and *political entity* were added. Widening the NE types to a hierarchy containing more than 200 types and subtypes (Sekine et al., 2002) enabled new perspectives for Question Answering systems and NER on data from social media like twitter (Ritter et al., 2011). NER systems may use grammar-based techniques as well as statistical models like machine learning. Systems using handcrafted rules obtain better precision by the price of lower recall and extensive linguistic work. Statistic systems require a large amount of expensive manually annotated training data. Recently, hybrid approaches were also explored to sidestep the drawbacks of both main techniques (Nothman et al., 2013). Often, gazetteer-based NER systems are also developed or integrated within already existing approaches (Jahangir et al., 2012). Current NER technologies still lack in performance in specific domains, such as politics, molecular biology or yellow press. For both rule-based and statistic systems, opportunities for new solutions are created (Poibeau and Kosseim, 2001). Furthermore, the identification of relevant expressions in text and automatically linking them to Wikipedia is part of the recent scope of NLP challenges (Mihalcea and Csomai, 2007). Additionally should be noted that NER systems for German are not easily available or are closed source.

3 Task Description

The main aim of the GermEval 2014 Named Entity Recognition Shared Task (Benikova et al., 2014a) is not only the detection of NEs, but as well the extension of the task specifically to one language – German. Additionally, GermEval increases the level of NE embedding, also targeting the identification of NEs inside already existing ones. Another peculiarity about the task is the fact that there are no restrictions regarding the types of NER systems as well as type and amount of used resources allowed for submission.

#	http://de.wikipedia.org/wiki/Manfred_Korfmai		
1	Aufgrund	O	O
2	seiner	O	O
3	Initiative	O	O
4	fand	O	O
5	2001/2002	O	O
6	in	O	O
7	Stuttgart	B-LOC	O
8	,	O	O
9	Braunschweig	B-LOC	O
10	und	O	O
11	Bonn	B-LOC	O
12	eine	O	O
13	große	O	O
14	und	O	O
15	publizistisch	O	O
16	vielbeachtete	O	O
17	Troia-Ausstellung	B-LOCpart	O
18	statt	O	O
19	,	O	O
20	„	O	O
21	Troia	B-OTH	B-LOC
22	-	I-OTH	O
23	Traum	I-OTH	O
24	und	I-OTH	O
25	Wirklichkeit	I-OTH	O
26	”	O	O
27	.	O	O

Figure 1: An example sentence of the GermEval data annotation format (Benikova et al., 2014b).

The data sets provided by the task consist mainly of articles extracted from the German Wikipedia and other News Corpora with over 31.000 sentences containing over 590.000 tokens. A sample of the data format can be seen in figure 1 (Benikova et al., 2014b). As the authors describe, the data is marked in the traditional BIO tagging scheme (Tjong Kim Sang and De Meulder, 2003) for the four main types: *person* (PER), *location* (LOC), *organization* (ORG) and *other* (OTH). Additionally, two subtypes with respect to all main classes are included: *part* and *deriv* indicating NE spans where only a subspan corresponds to a NE of the main types and respectively derivatives where the span is a derivation of a NE.

http://de.wikipedia.org

4 BECREATIVE

BECREATIVE is a Python implementation that makes use of the natural language toolkit (NLTK) that provides easy string handling, regular expression support and short development time. The current section provides further details about the system pipeline starting with preprocessing (see section 4.1), classification model (presented in section 4.2) and postprocessing (see section 4.3).

4.1 Preprocessing

During preprocessing, we bring the provided data, which is in a tab-separated value form, in a format that is better suited for our purpose. Internally we created a class representation for tokens, that mirrors the format of one row in the provided files and some empty fields for the tagger output, and one for sentences which is basically a list of tokens with some handy methods in addition. During the import, the data is already transformed to our representation of it, afterwards the data is annotated for part-of-speech (POS) by the TreeTagger developed by Helmut Schmid (Schmid, 1994; Schmid, 1995).

4.2 Naive Bayesian Classification

For NER proper, we train a Naive Bayesian classifier. The feature set used by the learner is presented in table 1. All feature representations are boolean values and the default weighting by the classifier is kept. The first 15 features are self-explanatory. Feature 16 checks if the second preceding token is a known NE (based on gazeteer lists collected from various online resources) and compares the preceding token against a list of verbs that indicate that the token could be a name. Feature 19 works similarly. Feature 17 checks the token for parts like *GmbH* or *Holding*, similar to 18 which tests for certain suffixes like *-hausen* or *ingen*. Feature 20 tests the second preceding token against a list of verbs, such as *wohnen* or *kommen* and looks the preceding token up in a list of prepositions.

4.3 Postprocessing

During postprocessing, gazetteer-based checks were additionally performed, which indicate a high probability of a token being a full or only part of a NE. The gazetteers were accumulated as lists for the following topics: Countries, Mountains, Waterbodies, Places of Interest, Street Names, Automobile Manufacturers, Book Titles, Film Titles, Styles, Forms of Address, First Names, Actors and Famous Persons.

As a final step, there is one list that contains phrases which are sure not to be Named Entities like measurements, so we are able to reduce the false positives a little further.

#	Description
1	The token itself
2	The preceding token
3	The following token
4	The token's index
5	The token's POS tag
6	The token's lemma
7	Capitalisation of the first letter
8	Capitalisation of the preceding word's first letter
9	Capitalisation of the following word's first letter
10	Whether the token matches a regular expression for a URL
11	Whether the token matches a regular expression for an IP address
12	Whether the token matches a regular expression for an email
13	Whether the token contains non letter characters
14	Whether the token contains numbers
15	Whether the token contains Roman numerals
16	Whether the token contextually could be a name
17	Whether the token has typical parts of an organization name
18	Whether the token has a location suffix
19	Whether the token contextually could be a location
20	Whether the token is one of certain verbs that stands usually with locations

Table 1: The feature set used by BECREATIVE

5 Evaluation

BECREATIVE was evaluated on the development set of the GermEval 2014 shared task. The results that the system achieves are presented in table 2. We also tested different subsets of the feature set. The first subset (*base*) includes features 1,2,3,4,7,8 and 9 from table 1, while the second subset (*base+POS*) adds the POS-tagger based features 5 and 6 as well. The performance of the full feature set is then listed under *all* in table 2. Additionally, after classification, the output of the classifier is also revised by our postprocessing gazetteer-based rules. leading to the system performance listed under *all+Lists* in table 2.

It is interesting to see (when the strict evaluation setting is observed) that including POS and lemma information in the feature set leads to a considerable decrease in system performance

http://www.python.org
http://www.nltk.org

setting	strict				loose				outer				inner			
	Acc.	P	R	F1	Acc.	P	R	F1	Acc.	P	R	F1	Acc.	P	R	F1
base	95.94	39.60	27.68	32.58	95.97	40.35	28.20	33.20	92.46	39.60	29.88	34.06	99.42	0.00	0.00	0.00
base+POS	92.75	18.07	42.21	25.31	92.83	19.06	44.51	26.69	86.07	18.07	45.58	25.88	99.42	0.00	0.00	0.00
all	95.90	38.66	31.89	34.95	95.93	39.29	32.42	35.52	92.38	38.66	34.44	36.43	99.42	0.00	0.00	0.00
all+Lists	95.97	39.58	35.34	37.34	95.99	40.20	35.90	37.93	92.51	39.58	38.16	38.86	99.42	0.00	0.00	0.00

Table 2: Results achieved by the BECREATIVE system based on the GermEval development set.

(from 32.58% for setting *base* to 25.31% for setting *base+POS*). This is due to the large decrease in precision (from 39.60% to 18.07%) even though recall is significantly improved (from 27.68% to 42.21%).

The combination of all features from table 2 leads to a system performance of 34.95% (see setting all), which is considerably low for a classification approach in comparison to state-of-the-art systems for German reported at the CoNLL-2003 Shared Task (Tjong Kim Sang and De Meulder, 2003).

Based on the setting for which all features are used ($all + Lists$), the detailed per class results given in table 3 show that BECREATIVE fails to identify most of the *part* and *deriv* subclasses (apart from LOCderiv and ORGpart). Additionally, all inner spans are also completely ignored by the system, which also contributes significantly to the overall low performance scores. This can be further approached by training two separate classifiers for both NE spans (outer and inner) and including span-specific or span-indicative features in both separate feature groups (e.g. classification decisions of the outer span can be included in the features for the inner span). Moreover, a task as NER would profit even more from sequential models (e.g. Conditional Random Fields) independent of the level of embedded phrases.

			P	R	F1
LOC	LOC	Outer strict	40.25	63.46	49.26
		Inner strict	0.00	0.00	0.00
		Outer loose	42.78	52.69	47.22
		Inner loose	0.00	0.00	0.00
	LOCderiv	Outer strict	63.95	23.91	34.81
		Inner strict	0.00	0.00	0.00
	LOCpart	Outer strict	0.00	0.00	0.00
		Inner strict	0.00	0.00	0.00
ORG	ORG	Outer strict	27.21	25.45	26.30
		Inner strict	0.00	0.00	0.00
		Outer loose	28.24	22.62	25.12
		Inner loose	0.00	0.00	0.00
	ORGderiv	Outer strict	0.00	0.00	0.00
		Inner strict	0.00	0.00	0.00
	ORGpart	Outer strict	37.50	3.30	6.06
		Inner strict	0.00	0.00	0.00
OTH	OTH	Outer strict	51.24	22.96	31.71
		Inner strict	0.00	0.00	0.00
		Outer loose	51.24	20.46	29.25
		Inner loose	0.00	0.00	0.00
	OTHderiv	Outer strict	0.00	0.00	0.00
		Inner strict	0.00	0.00	0.00
	OTHpart	Outer strict	0.00	0.00	0.00
		Inner strict	0.00	0.00	0.00
PER	PER	Outer strict	41.65	40.65	41.15
		Inner strict	0.00	0.00	0.00
		Outer loose	41.65	39.53	40.57
		Inner loose	0.00	0.00	0.00
	PERderiv	Outer strict	0.00	0.00	0.00
		Inner strict	0.00	0.00	0.00
	PERpart	Outer strict	0.00	0.00	0.00
		Inner strict	0.00	0.00	0.00

Table 3: Results per class achieved by BECREATIVE based on the GermEval development set.

6 Future Work and Conclusion

The current paper presented the BECREATIVE system for NER developed and evaluated in the context of the GermEval 2014 Named Entity Recognition Shared Task. BECREATIVE combines a Naive Bayesian Classifier with rules performing gazetteer-based checkup and achieves a performance of 37.34 on the development set.

In the future, we plan to explore further features (e.g. investigating for example a larger context than just preceeding and following tokens) for the classification approach in order to improve the still considerably low learner performance. Additionally, as noted above, we would also like to apply sequential models to the task and include a separate classification for each layer of embedding present in the data.

References

Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Pado. 2014a. Germeval 2014 named entity recognition: Companion paper. In *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition*, Hildesheim, Germany.

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014b. NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Xavier Carreras, Lluís Màrquez, and Lluís Padró. 2002. Named entity extraction using adaboost. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: A brief history. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, COLING '96, pages 466–471, Stroudsburg, PA, USA. Association for Computational Linguistics.

Faryal Jahangir, Anwar Waqas, Bajwa Usama Ijaz, and Wang Xuan. 2012. N-gram and gazetteer list based named entity recognition for urdu: A scarce resourced language. In *Proceedings of the 10th Workshop on Asian Language Resources*, page 95–104.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *In CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242. ACM.

1995. *MUC6 '95: Proceedings of the 6th Conference on Message Understanding*, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artif. Intell.*, 194:151–175.

Thierry Poibeau and Leila Kosseim. 2001. Proper name extraction from non-journalistic texts. In *In Computational Linguistics in the Netherlands*, pages 144–157.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA. Association for Computational Linguistics.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Helmut Schmid. 1995. Improvements In Part-of-Speech Tagging With an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*, pages 47–50, Dublin, Ireland.

Satoshi Sekine and Hitoshi Isahara. 2000. Irex: Ir and ie evaluation project in japanese. In *Proceedings of International Conference on Language Resources & Evaluation (LREC 2000)*.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended named entity hierarchy.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 Shared Task: Language-independent Named Entity Recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.